利用专家指导的多智能体强化学习提升短视频流媒体的自适应比特率性能

Yueheng Li*南京大学 中国

Qianyuan Zheng*
南京大学 中国

Zicheng Zhang 南京大学 中国

Hao Chen†
南京大学 中国

Zhan Ma 南京大学 中国

arXiv:2304.04637v1 [cs.MM] 10 Apr 2023

摘要

在短视频流媒体领域，为传统长视频应用开发的热门自适应比特率（ABR）算法，由于仅针对比特率自适应进行调优，因而会遭遇灾难性失效。相比之下，短视频自适应比特率（SABR）算法必须协同确定预取内容的视频及其比特率等级，同时既不牺牲用户的体验质量（QoE），也不造成明显的带宽浪费。遗憾的是，现有的 SABR 方法不可避免地受到收敛缓慢和泛化能力差的困扰。因此，本文提出了一种新型 SABR 框架 Incendio，该框架采用带有专家指导的多智能体强化学习（MARL），将视频 ID 和视频比特率的决策分别交由缓冲区管理智能体和比特率自适应智能体负责，从而最大化系统层面利用率得分；该得分建模为体验质量和带宽浪费指标的复合函数。训练 Incendio 时，首先通过模仿人工设计的专家规则进行初始化，随后再利用 MARL 进行微调。大量实验结果表明，Incendio 的表现优于当前最先进的 SABR 算法，效用得分提升了 53.2%，同时保持较低的训练复杂度和推理时间。

CCS 概念

• **信息系统** →**多媒体流媒体**。

关键词

短视频流媒体、自适应比特率、强化学习、模仿学习

ACM 参考格式：

YuehengLi、Qianyuan Zheng、ZichengZhang、Hao Chen 和 Zhan Ma。2023年。使用专家指导的多智能体强化学习提升短视频流媒体的自适应比特率性能。载于会议论文集。

*两位作者对本研究的贡献相同。

†Hao Chen是通讯作者，*chenhao1210@nju.edu.cn*。

允许出于个人或课堂用途免费制作本作品全部或部分内容的数字或纸质副本，前提是这些副本不得用于营利或商业利益，且每份副本的首页均须附有本声明和完整引文。必须尊重本作品中归他人而非 ACM 所有的组件的版权。注明出处后可以摘录。以其他方式复制、重新发布、发布至服务器或重新分发至列表，均需事先获得明确许可和/或支付费用。请向 permissions@acm.org 申请许可。

*NOSSDAV '23，2023*年*6*月*7*日至*10*日，加拿大温哥华© 2023 计算机协会。
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

第*33*届数字音频与视频网络及操作系统支持研讨会（*NOSSDAV'23*）。ACM，美国纽约州纽约市，7页。 https://doi.org/XXXXXXX.XXXXXXX

1 引言

近年来，使用 Kwai [7] 和 TikTok [14]等短视频流媒体应用进行娱乐、社交联系等活动的情况大幅增加，导致流量呈指数级增长。这类短视频服务与传统的长视频流媒体场景（如视频点播（VoD））有很大不同，因为用户只需滚动触摸屏，就能迅速切换到自己感兴趣的内容。为此，我们通常需要将个性化内容适当地预取到本地缓冲区中。尽可能多地将短视频预取到本地缓冲区，可以确保消费过程中的体验质量（QoE），但往往会导致严重的带宽浪费。另一方面，缓冲不足可能会在滚动到下一个尚未缓存的视频时造成明显的启动延迟。

为应对这些挑战，内容提供商主要采用短视频自适应比特率算法（SABR），以同时优化用户的体验质量并减少带宽浪费。为此，SABR 算法需要确定要下载哪个视频或保持空闲，然后确定要预加载该特定视频的哪种比特率。为实现这一目标，可以综合利用网络状况、客户端缓冲区状态、分块大小以及用户的观看偏好，从而做出适当的决策。ACM Multimedia2022（MMGC2022 [22]）举办了一项短视频流媒体大型挑战赛，吸引了众多具有竞争力的解决方案。

研究人员开发了基于规则和基于强化学习（RL）的SABR方法。例如，PDAS [21], 作为一种典型的基于规则的方法，在MMGC2022中展现出领先的性能；该方法采用基于概率的奖励函数和手工设计的缓冲区管理模型。然而，基于规则的方法常因难以推广到不同环境而受到批评，因为固定的控制规则无法在实践中全面刻画所有场景下的系统行为 [1, 8]。此外，由于PDAS是模型预测控制（MPC [15]）的一种变体，采用贪心启发式搜索进行决策，因此随着优化时间范围长度的增加，其决策推理时间呈指数增长。

因此，人们引入了基于强化学习的方法，通过神经网络直接建立环境观测与动作之间的联系，以克服这些问题。例如，DAM [9],是一种基于强化学习的 SABR 方法 [3, 13],，它根据各原子动作组合的概率，同时做出缓冲区管理和比特率自适应决策

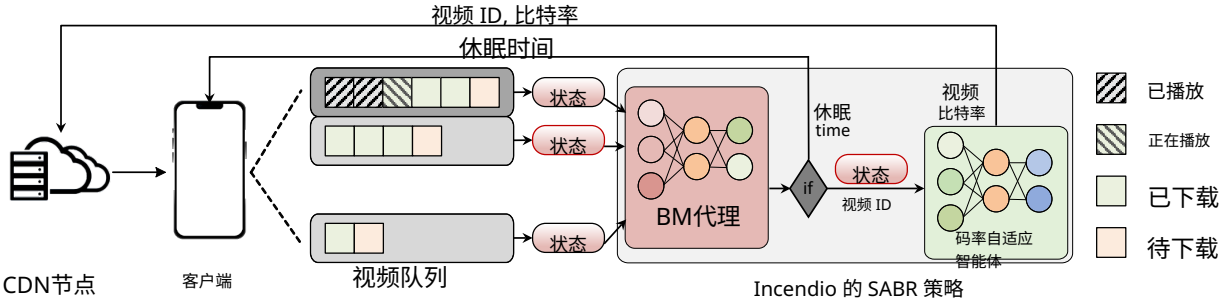

图 1：*Incendio* 使用两个分层智能体，分别负责缓冲区管理（BM 智能体）和码率自适应（BA 智能体）。它们在每次决策迭代中，根据包括过去的吞吐量测量值、用户留存率、视频分块大小和缓冲区状态在内的观测值做出决策。

。在 DAM 的训练过程中，由于需要在庞大的离散动作空间中进行搜索，而该空间与队列中的视频数量以及每个视频的总码率级别密切相关，因此 DAM 的收敛速度较慢（因而耗时极长），难以达到全局最优性。

因此，本文提出了Incendio，这又是一种新颖的 SABR 框架，旨在解决现有方法在体验质量与带宽效率联合优化方面的上述问题。我们以顺序方式分别对缓冲区管理和码率自适应的决策进行处理，即进行子任务分解，并据此设计分层多智能体强化学习（多智能体强化学习），同时训练这两个子任务，以优化复合奖励。这大幅缩小了最优性搜索的动作空间，从而以更快的收敛速度加速神经网络训练。另一方面，我们不再从头开始执行多智能体强化学习，而是提出利用人类经验，通过模仿学习将 Incendio从初级状态预训练至专家状态，这进一步减少了多智能体强化学习中的无效试验次数，也降低了陷入次优解的风险。

我们在各种网络和用户偏好条件下（详见 §5.1），将 Incendio 与 PDAS [21]、MPC [15]，和 DAM [9]，等先进算法进行比较，以评估 Incendio 的性能。结果表明，在所有场景中，Incendio 的表现始终优于现有算法。按总体效用得分衡量（如 §5.2 所述），Incendio 的平均表现比获奖算法 PDAS 高出 53.2%，同时保持了出色的训练效率（如 §5.3 所述）和部署可行性（如 §5.4 所述）。

2 背景与相关工作

本节首先简要介绍 SABR 问题中广泛认可的优化目标函数。随后，我们回顾基于规则和基于学习的 SABR 算法，并讨论其局限性。

优化目标。与主要关注提升用户体验质量（QoE）的长视频流媒体应用（例如视频点播）不同，短视频流媒体需要同时考虑体验质量提升和带宽效率（例如减少带宽浪费）。因此，SABR 的优化目标不仅涉及 QoE 模型 [1, 8, 15]，还包括带宽使用惩罚项，其定义为总体效用得分 [22],，即：

$$U_i = QoE_i - Bandwidth_i = \sum_m (R_m - S_m) - \sum_n \mu \cdot T_n - \sum_n v \cdot bw_n, \quad (1)$$

其中，m and n 分别表示视频中已播放和已下载的视频块数量 i。R_m 和 S_m 分别表示每个已播放视频块 m的质量（比特率）及其波动。T_n 和bw_n 分别表示下载视频块 n导致的重新缓冲时间和带宽使用量。我们按照 [22] 中的建议设置系数$\mu = 1.85$, $v = 0.5$，以确保与其他方法进行公平比较。

基于规则的 SABR 方法。APL [18] 通过使用李雅普诺夫优化提出了一种自适应预加载机制，以同时最大化播放流畅度并尽量减少带宽浪费。然而，APL 对短视频作出了固定比特率的假设，这在实际应用中并不现实。PDAS [21]纳入了用户留存率[1]，以更准确地预测体验质量，并利用模型预测控制规则，通过比较未来操作的所有可能组合来辅助决策，从而取得了最先进的性能（在 MMGC2022 中排名第一）。不过，PDAS 的超参数高度依赖具体情境，导致该模型难以泛化到用户偏好和网络状况各异的环境中（详见 §5.2）。

基于学习的方法 已在传统的自适应比特率任务中展现出优越性 [5, 8, 15]。对于 SABR，LiveClip [4] 采用强化学习来预测视频切换事件并动态调整预加载顺序，但忽略了比特率自适应和带宽节省问题。DUASVS [17]利用集成学习，为预取阈值和视频比特率的决策制定控制策略。DAM [9]通过将用户留存率纳入奖励函数，并利用动作掩码降低训练复杂度，实现了卓越的性能（在 MMGC2022 的所有基于学习的技术中排名第一）。然而，鉴于 SABR 任务中的离散动作空间较大，而该空间源于是否休眠、

[1]用户留存率通过统计选择继续观看当前视频的用户比例来体现，内容提供商可以以分块为粒度提供该数据。

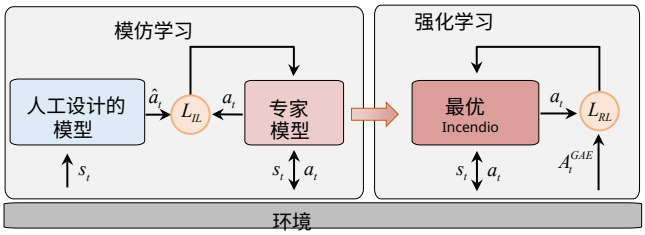

图 2：Incendio 的两阶段训练：首先通过模仿手工设计的模型进行初始化，然后通过强化学习进行微调。

待预取的视频 ID 以及比特率等级等多维决策（详见 §5.3），上述基于学习的方法在训练过程中仍存在收敛速度缓慢的问题。

3 系统概述

Incendio 的系统架构如图 1 所示。每段短视频被切分为长度为 1 秒的视频块。每个视频块都编码为多个比特率版本，并存储在内容分发网络（CDN）节点中。客户端会从 CDN 节点主动下载视频块，并为视频队列中的每段短视频维护一个本地缓冲区，其中包括当前正在播放的视频和几段推荐视频。不同视频以不同 ID 标记。每当用户滚动屏幕时，队列中的第二个视频便开始播放，而前一个视频中已下载但尚未播放的视频块会被清除，造成带宽浪费。与此同时，视频推荐机制推荐的新视频会被添加到队列中。

Incendio 的强化学习智能体由两个分层智能体组成，分别负责缓冲区管理（BM 智能体）和码率自适应（BA 智能体）。在每次决策迭代中，BM 智能体根据过去吞吐量测量值的观测值，以及队列中每段视频的状态（包括用户留存率、剩余缓冲区大小、平均视频块大小、重缓冲时间、比特率及最近一次已下载视频块的比特率波动），选择休眠一段固定时间，或选择最需要缓冲的视频。如果 BM 智能体决定休眠，客户端的预加载过程将暂停预先设定的时长。否则，BM 智能体会根据视频 ID 决定预取哪段视频。随后，BA 智能体根据该视频的状态和网络状况，确定要下载的下一个视频块的比特率。之后，客户端向 CDN 节点提交包含视频 ID 和比特率的请求，开启新一轮交互。

视频的数量以及队列中每个视频的码率级别由底层流媒体平台决定。这里我们采用MMGC2022中使用的相同设置，即队列中包含五个视频，且每个视频有三个码率级别。离线训练完成后，Incendio 策略在任务推理过程中保持固定。下一节将详细介绍 Incendio 的智能体设计和训练方法。

4 INCENDIO 设计

如图 2 所示，Incendio 的训练过程分为两个阶段。第一阶段，我们分别通过模仿人工设计的专家策略，对 Incendio 的两个智能体进行预训练，从而避免它们在训练初期进行大量低效探索。随后，使用多智能体强化学习（MARL）进一步微调 Incendio 的策略，直至收敛到全局最优解。值得注意的是，Incendio 采用集中训练、分散执行（CTDE）的方法训练其两个智能体，使它们协作以实现共同的奖励目标。这种训练策略不仅有助于高效更新 Incendio 的策略，还能避免其收敛到次优策略。

本节首先介绍Incendio 智能体的输入状态、动作和神经网络，这些内容在预训练和微调训练阶段保持一致。随后详细阐述预训练和微调的训练算法。

4.1 多智能体设计

Incendio的多智能体将客户端收集的观测值作为输入状态 s_t，根据其由神经网络表示的策略 $\pi_\theta(s_t, a_t)$采取行动 a_t。本小节详细阐述 Incendio 的状态、行动和神经网络设计。

状态。 Incendio 在步骤 t 收集的状态定义为$s_t = (\vec{b_t}, \vec{l_j}, \vec{g_j}, \vec{u_j}, \vec{h_j}, \vec{q_j}, \vec{f_j})$。第一个分量是过去 K个视频块的吞吐量测量值向量（例如，$\vec{b_t} = \{b_{t-K+1}, \ldots\ldots, b_t\}$），每个值都可以通过将视频块大小除以下载时长来计算。其余状态分量表示队列中每个视频 $j \in [1,5]$ 的不同类型因素。具体来说，l_j 是用户从当前正在播放的视频块 j m（对于推荐视频，$m = 1$）继续观看该视频直至视频块 n的条件概率，可按如下方式计算：

$$l_j = \frac{p_j^n}{p_j^m}, \tag{2}$$

其中，p_j^m表示通过统计平均得到的视频 j 在视频块 m 的用户留存率；g_j 表示视频j当前的缓冲区大小；u_j 表示视频 j后续视频块在不同比特率下的平均大小；h_j 表示下载视频 j 的上一个视频块所导致的重新缓冲时间，若未发生重新缓冲则为 0；q_j 和 f_j 分别表示视频 j上一个视频块已下载时的比特率及其波动。比特率波动可通过以下方式获得

$$f_j = \left|q_j - q_{j-1}\right|. \tag{3}$$

BM 智能体的状态定义为 $s_t^{bm} = (\vec{b_t}, \vec{l_j}, \vec{g_j}, \vec{u_j})$。BA 智能体针对视频 j做出决策时的状态定义为$s_t^{ba} = (\vec{b_t}, l_j, g_j, u_j, h_j, q_j, f_j)$。在本研究中，我们根据经验将 $K = 5$设定为，以捕获过去观测值中的时间特征。

动作。 BM 智能体的输出动作 a_t^{bm}是一个长度为 6 的 0-1 向量。该向量的前五个值表示相应的视频 ID，最后一个值表示休眠固定时长 $\tau = 200ms$。这一设置旨在平衡计算资源的最优利用与不错过理想决策时间之间的权衡。类似地，BA 智能体从 {750，1200，1850} kbps 的比特率梯度中选择一个动作 a_t^{ba}（与 MMGC2022 相同），这些比特率对应不同的视频质量。

神经网络。 Incendio 使用演员-评论家框架来表示其控制策略。BM 智能体的演员网络结构如图 3 所示。为捕捉

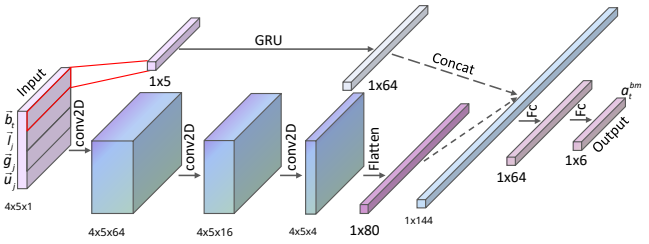

图 3：Incendio 的 BM 智能体的神经网络架构。

过去带宽的时间特征，我们采用包含 64 个单元的门控循环单元（GRU）层。此外，为提取输入向量之间的空间特征，我们使用了三个二维卷积网络层，卷积核大小为 5x5，输出通道数分别为 64、16 和 4。随后将输出拼接起来，并传入包含 64 个单元的全连接层。网络的每一层都使用 Leaky-ReLU 作为激活函数。在输出层，我们在全连接网络中使用 Soft-Max 激活函数，得到一个 6 维向量，用于表示选择各个动作的概率。对于 BA 智能体的演员网络，我们用一个全连接层替换卷积层，其余部分与 BM 智能体保持一致。BM 智能体和 BA 智能体共用一个评论家网络，该网络将 BM 智能体演员网络的最后一个隐藏层与 BA 智能体演员网络的最后一个隐藏层拼接起来，输出一个不经过激活函数的张量作为值。这个精心设计的神经网络使 Incendio 能够从观测值中高效提取时空特征，最终取得出色的性能。

4.2 使用模仿学习进行预训练

模仿学习是一种机器学习方法，智能体通过观察专家生成的轨迹来学习执行任务，并已被证明在机器人、自动驾驶汽车和网络流媒体等多个领域中十分有效 [5, 6, 20]。我们使用模仿学习对 Incendio 进行预训练，是因为它可以减少智能体学习任务所需的试错次数，从而实现更快的收敛和更高效的学习。

专家策略。 专家策略的质量直接决定了智能体通过模仿学习所能达到的上限。受近期一项将基于学习的方法与基于规则的方法相结合以进一步提升性能的研究启发 [19]，我们采用先进的手工设计方法 PDAS [21] 来指导 Incendio。如 §2 所述，PDAS 将用户留存率纳入考量，以提高体验质量估算的准确性，并采用模型预测控制技术，通过穷尽分析所有可能的未来决策组合，实现最优决策过程，从而取得良好性能。在本工作中，我们使用 PDAS 设计的模型，根据未来一定时间范围内视频块的实际吞吐量，估算每种可能动作组合的体验质量，并选取体验质量最高的组合作为专家策略。随后，我们将该专家策略拆分为两个子集：视频 ID 轨迹和比特率轨迹，分别用于训练 Incendio 的两个智能体。更多详情请参阅 PDAS。需要注意的是，PDAS 主要用于提高训练初期探索的效率，也可以替换为其他优秀的基于规则的方法。

损失函数。 与传统监督学习中的样本由特征—标签对组成类似，模仿学习以状态—动作对的示范为特征。因此，广泛用于分类问题的交叉熵函数也适用于此处。Incendio 模仿学习的损失函数如下所述：

$$L_{IL} = -\sum_{t} \hat{A}_t \log \pi_\theta(s_t, a_t), \tag{4}$$

其中，$\pi_\theta(s,\ a)$ 是参数为 θ 的智能体策略。$\hat{}A$ 是由专家策略生成的动作概率列表，其中专家动作 $\hat{}a$ 的值等于 1，其余动作的值均为 0。我们将策略梯度方法 [12] 作为 IL 和 RL 的训练策略。其主要思想是估计期望总奖励相对于策略参数 θ 的梯度，并根据这些梯度更新网络参数，具体可写为：

$$d\theta \leftarrow d\theta - \alpha \sum_{t} \nabla_\theta \hat{A}_t \log \pi_\theta(s_t, a_t). \tag{5}$$

我们设置学习率 $\alpha = 0.0001$。Incendio 的 BM 智能体和 BA 智能体使用相同的损失函数进行训练。我们还引入了一种新颖的经验回放机制，以提高样本利用效率。

4.3 使用强化学习进行微调

Incendio 使用先进的多智能体强化学习框架微调其 SABR 策略，即多智能体近端策略优化（MAPPO [16]）算法。该算法是近端策略优化（PPO [11]）针对多动作任务设计的一种有效改进版本。在每个交互步骤，Incendio 根据其策略 π_θ 和观测值 s_t 执行动作 a_t，其中包括 a_t^{bm} 和 a_t^{ba}。随后，环境转移到新的状态 s_{t+1} 并返回奖励 r_t，用于更新 Incendio 两个智能体的神经网络参数。从一个智能体的视角来看，另一个智能体带来的变化会被视为环境变化。

奖励。 在典型的强化学习任务中，智能体通过最大化从环境中获得的期望累积（折扣）奖励来学习最优策略。因此，我们将奖励设置为反映公式（1）中定义的 MMGC2022 效用得分。为了最大化期望累积折扣奖励，Incendio 的强化学习智能体学习获得更高的效用得分。为了促使Incendio学习带宽效率更高的策略，我们在奖励函数的第一项中添加系数 w_n，其可以表述如下：

$$r_t = w_n \cdot l_n \cdot (R_n - S_n) - \mu \cdot T_n - \nu \cdot bw_n, \tag{6}$$

其中，l_n 是方程（2）中定义的条件概率。其背后的理念是，如果下一个待下载分块的用户留存率远低于当前正在播放分块的用户留存率，这意味着用户很可能会在播放下一个分块之前滑走，智能体将获得较低的奖励。为了获得更高的奖励，智能体会尝试暂停下载，直到条件概率增加，从而减少带宽浪费。w_n 表示用户是否会观看最后一个已下载分块 n，其定义为

如下：

$$w_n = \begin{cases} 0, & \text{if user will scroll to next video before chunk } n \\ 1, & \text{if user will watch chunk } n. \end{cases} \quad (7)$$

在训练过程中，每当一个视频被添加到视频队列时，w_n 会在用户留存率轨迹中被随机采样。在评估期间，我们仍使用公式（1）来评估性能。

训练方法。我们使用裁剪替代损失函数训练 Incendio 的强化学习智能体，并额外引入策略熵，以避免在训练早期收敛到次优策略。损失函数表述如下：

$$L_{RL} = -\sum_t \sum_k \min\left(ratio_{\theta,t}^k, \text{clip}\left(ratio_{\theta,t}^k, 1-\epsilon, 1+\epsilon\right)\right) A_t^{GAE} + \beta \sum_t \sum_k H\left(\pi_\theta\left(s_t^{(k)}\right)\right). \quad (8)$$

这里 A_t^{GAE}是使用 GAE [10]方法计算的优势函数，它表示智能体在状态 s_t中确定性地选择动作 a_t 时的预期奖励，与遵循具有策略参数 θ的策略 π_θ选择动作时的预期奖励之差。$ratio_{\theta,\ t}^k$ 和 $H(\cdot)$分别表示替代目标和策略熵。k 对智能体进行索引，其中 $k \in \{bm,\ ba\}$。 β是熵项的权重，当奖励在 100个训练轮次内没有增加时，我们会对其进行衰减。梯度更新公式与公式（5）类似，有关训练算法的更多技术细节请参见 [16]。

5 评估

5.1 方法论

为了评估 Incendio 的性能，我们使用 MMGC2022 提供的多视频模拟器，通过随机组合不同的视频轨迹和网络轨迹，模拟各种短视频流媒体会话。对于视频轨迹，不同比特率下每个分块的大小记录在视频大小轨迹中，而每个分块对应的用户留存率则记录在用户留存率轨迹中。

为了训练 Incendio，我们通过组合包括Oboe [1] 和 FCC [2] 在内的一些公开数据集，创建了网络轨迹语料库。对于视频轨迹，我们使用 DUASVS [17]，其中包含数百万条记录，包括视频分块统计数据和用户留存率。除非另有说明，否则我们从语料库中随机抽取 80% 作为 Incendio 的训练集，其余 20% 作为所有 SABR 算法的测试集。为进行公平比较，同一训练集也用于训练 DAM。由于MMGC2022中的网络轨迹和视频轨迹都相对较少，我们仅使用该 MMGC2022 数据集进行评估。请注意，byMMGC2022提供的网络轨迹分别记录了高、中、低网络条件下随时间变化的带宽采样。

我们将 Incendio 与以下方案进行比较，这些方案共同代表了当前最先进的方法：

- 模型预测控制*+fix-preload*，将一种常见的基于模型的自适应比特率方法（即 RobustMPC [15]）与 *fix-preload* [22]，中采用的休眠机制相结合；后者是 MMGC2022 提供的基线。RobustMPC 最大化由以下内容定义的累积效用得分：基于缓冲区占用观测和吞吐量预测，在未来 5 个视频块的时间范围内使用公式 (1)。
- *DAM* [9], 一种基于深度强化学习的方法，用于训练控制策略以做出视频 ID 的决策，bi - 码率等级、暂停/休眠时间，并结合动作掩码机制。它在 MMGC2022 的所有基于学习的技术中排名第一。我们严格按照其论文实现了 DAM。
- *PDAS* [21], MMGC2022 中的最先进方法，该方法通过联合利用用户留存率和手工设计的缓冲区管理模型来优化效用得分。PDAS 采用 RobustMPC 以实现最优决策。我们根据其论文忠实地实现了 PDAS。

5.2 总体结果

图4和图5展示了 Incendio 与其他方案在平均效用得分指标及其完整累积分布函数（CDF）方面的性能比较结果，实验采用了不同的短视频数据集（MMGC2022 和 DUASVS）及带宽数据集（MMGC2022 和 Oboe/FCC）。如图4(a)和图5(a)所示，Incendio 在平均得分及若干单项指标（即重新缓冲、流畅度和带宽浪费）上均明显领先。具体而言，Incendio 的平均效用得分比最先进的 PDAS 高出 53.2%。此外，Incendio 将重新缓冲减少了 39.1%–61.8%，将带宽浪费减少了 34.9%–59.2%，相较于其他方案取得了显著提升。图4(b)和图5(b)所示的完整累积分布函数结果进一步证明了 Incendio 性能的一致性。

如图4(a)所示，PDAS+fixed- preload 在几乎所有指标上都优于模型预测控制，表明它是模型预测控制的一个更先进的变体。然而，PDAS 在带宽节省和控制比特率波动方面的表现不如 DAM，这说明 PDAS 的最大缓冲区模型设计得不够完善。另一方面，在 Oboe/FCC 网络条件下使用 DUASVS 视频数据集时，PDAS 的表现最差，这揭示了其泛化能力较弱。我们认为，这是因为在 MMGC2022 数据集上调优的 PDAS 超参数无法刻画 Oboe/FCC 轨迹中的网络动态，也无法反映 DUASVS 视频集中用户偏好的变化。

有趣的是，我们发现，正如图4(b)和图5(b)所示，DAM 在不同 QoE 区间的表现不稳定：其在高 QoE 区间（约为 > 0.8）表现相对较好，但与其他竞争方案相比，在低和中等 QoE 区间（< 0.8）的表现明显较差。原因在于，面对庞大的多维探索空间，DAM 倾向于舍弃某些动作，以便促进探索过程。此外，我们从 DAM 的训练日志中发现了支持上述分析的证据：对于推荐的视频，它总是选择中等比特率，从不为当前正在播放的视频选择最低比特率。DAM 的指标表现也与我们的发现一致，这在一定程度上是因为选择高质量（高比特率）视频块会增加重新缓冲的概率。

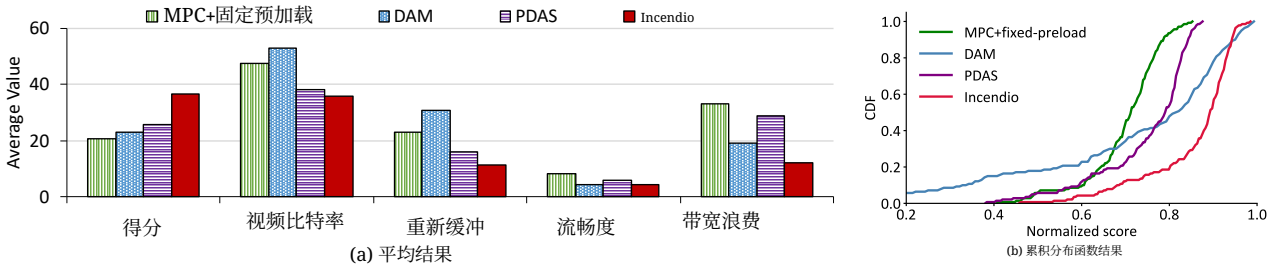

图4：在 MMGC2022视频数据集和带宽轨迹数据集上，比较 Incendio 与其他方案的平均性能和完整累积分布函数性能。累积分布函数结果中的得分已归一化。

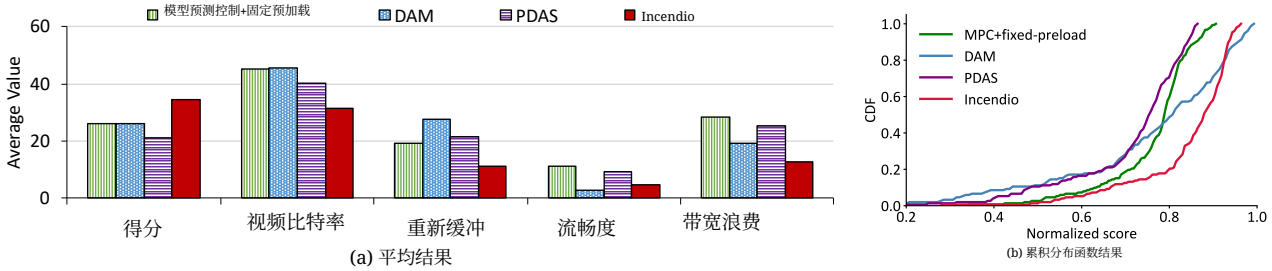

图 5：在 DUASVS 视频数据集和 Oboe/FCC 带宽跟踪数据集下，从平均性能和完整累积分布函数性能方面比较 Incendio 与其他方案。累积分布函数结果中的效用得分已进行归一化。

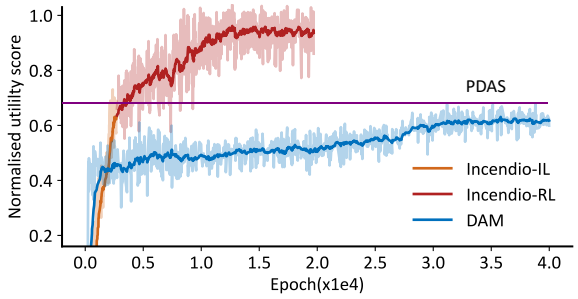

图 6：Incendio 和 DAM 的训练日志。

5.3 训练效率

我们在图 6 中绘制了 Incendio 和 DAM 的训练日志，以比较我们的两阶段训练方法与 DAM的集中式强化学习方法的训练效率。如图所示，DAM 在训练早期陷入了次优策略，并且需要很长的收敛时间才能达到最终策略。这是因为所有可能的原子动作组合所产生的巨大探索空间给 DAM 带来了重大挑战。相比之下，Incendio 展现出了通过模仿学习在显著更短的 4k训练轮次内学习专家级策略的能力。此外，我们提出的两阶段训练框架使Incendio 能够在强化学习阶段取得显著进展。这些结果证明了我们训练方法的效率和重要性。

5.4 推理时间

为了评估 Incendio 的复杂度，我们记录其推理阶段的运行时间，并将其与使用 MPC+fix-preload、DAM 和 PDAS 时的运行时间进行比较。我们设置了两种不同的环境进行评估。在环境 $E1$中，我们保持与……相同的设置。

表1：Incendio 和现有SABR 算法在两种不同环境中的推理时间。时间以毫秒为单位测量。

	模型预测控制+固定预加载	DAM	PDAS	**Incendio**
E1	0.6	0.4	7.1	**1.0**
E2	11.8	0.4	318.1	**1.1**

MMGC2022竞赛中，视频和码率级别的数量分别略微增加至 7 和 6。实验在配备 Intel(R)Core(TM) i5-12500@3.00GHzCPU 的台式机上进行，并重复数千次。表 1 列出了 MPC+fix-preload 、 DAM 、 PDAS 和 Incendio 的平均推理时间。如表所示，随着队列中的视频数量和码率级别增加，MPC+fix-preload 和PDAS 的计算复杂度呈指数增长，这对于移动设备而言是不可接受的。这是因为这两种基于 MPC的 SABR算法通过比较未来动作所有可能组合的期望奖励来做出决策，这限制了其在动作空间较大的场景中的应用。相反，Incendio在这些环境中保持了较低的计算复杂度，其计算复杂度与DAM处于同一数量级，但提供了更优的 SABR 性能（§5.2）和更高的训练效率（§5.3）。

6 结论

我们提出并评估了 Incendio，这是一种面向短视频流媒体的新型自适应比特率框架，采用专家指导的多智能体强化学习。该框架以顺序方式将缓冲区管理（例如视频预取）与比特率自适应的决策分离，以优化系统级效用分数。因此，Incendio 的训练首先通过模仿学习采用人类专家策略，随后使用小得多的动作空间

进行策略微调，从而显著提升收敛速度，并展现出卓越的泛化能力。在广泛的测试条件下，我们发现，采用效用得分衡量时，所提出的 Incendio 相较于当前最先进的方案，性能提升超过 2 倍。

致谢

本研究部分受到中国国家自然科学基金（62101241）和江苏省双创博士项目（JSSCBS20210001）的资助。

参考文献

[1] Zahaib Akhtar、Yun Seong Nam、Ramesh Govindan、Sanjay Rao、Jessica Chen、Ethan Katz-Bassett、Bruno Ribeiro、Jibin Zhan 和 Hui Zhang。2018。Oboe：根据网络状况自动调优视频自适应比特率算法。载于《2018 年计算机协会数据通信特别兴趣小组会议论文集》（匈牙利布达佩斯）（SIGCOMM '18）。计算机协会，美国纽约州纽约市，44 58 10 1145 3230543 3230558 –。[2] 2016。https://doi.org/。/。美国联邦通信委员会。。原始数据——《测量美国宽带 2016》。检索于 2022 年 3 月 1 日，网址：https://www.fcc.gov/reports-research/reports/measuring-broadband-america/raw-data-measuring-broadband-america-2016[3] Ting-Han Fan 和 Yubo Wang。2022。具有整数动作的软演员-评论家算法。载于《2022 年美国控制会议》（ACC）。IEEE，2611–2616。[4] Jianchao He、Miao Hu、Yipeng Zhou 和 Di Wu。2020。LiveClip：利用深度强化学习实现智能移动短视频流媒体。载于《第 30 届计算机协会网络与操作系统对数字音频和视频支持研讨会论文集》（土耳其伊斯坦布尔）（NOSSDAV '20）。计算机协会，美国纽约州纽约市，54–59。https://doi.org/ 10 1145 3386290 3396937 。/。Tianchi Huang、Chao Zhou、Rui-Xiao Zhang、Chenglei Wu、[5] Xin Yao 和 Lifeng Sun。2019。Comyco：通过模仿学习实现质量感知型自适应视频流媒体。载于《第 27 届 ACM 国际多媒体会议论文集》（法国尼斯）（MM '19）。计算机协会，美国纽约州纽约市，429 437 10 1145 3343031 3351014 –。https://doi.org/。/。Ahmed Hussein、Mohamed [6] Medhat Gaber、Eyad Elyan 和 Chrisina Jayne。2017。模仿学习：学习方法综述。《ACM 计算调查》50 卷第 2 期，文章 21 2017 35 10 1145 3054912（4 月），页数。https://doi.org/。/。快手。。快手。[7] 2022。检索于 6 月 16, 2022，网址：https://www.kwai.com[8] Hongzi Mao、Ravi Netravali 和 Mohammad Alizadeh。2017。利用 Pensieve 实现神经自适应视频流媒体。载于《计算机协会数据通信特别兴趣小组会议论文集》（美国加利福尼亚州洛杉矶）（SIGCOMM '17）。计算机协会，美国纽约州纽约市，197 210 10 1145 3098822 3098843 –。https://doi.org/。/。Si-Ze Qian、Yuhong Xie、Zipeng [9] Pan、Yuan Zhang 和 Tao Lin。2022。DAM:基于深度强化学习并采用动作屏蔽的短视频流媒体预加载算法。载于《第 30 届 ACM 国际多媒体会议论文集》（葡萄牙里斯本）（MM '22）。计算机协会，美国纽约州纽约市，7030 7034 10 1145 3503161 3551573 –。https://doi.org/。/。John [10] Schulman、Philipp Moritz、Sergey Levine、Michael Jordan 和 Pieter Abbeel。2015。使用广义优势估计进行高维连续控制。https://doi.org/ 10 48550 1506 02438 。/ARXIV.。John [11] Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。2017。近端策略优化算法。https://doi.org/10.48550/ARXIV. 1707.06347[12] Richard S Sutton、David McAllester、Satinder Singh 和 Yishay Mansour。1999。函数逼近下的强化学习策略梯度方法。载于《神经信息处理系统进展》，S. Solla、T. Leen 和 K. Müller（编），第 12 卷。MIT 出版社。https://proceedings.neurips.cc/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf[13] Yunhao Tang 和 Shipra Agrawal。2020。将连续动作空间离散化以进行同策略优化。载于《AAAI 人工智能会议论文集》，第 34 卷。5981–5988。[14] TikTok。2022。TikTok。检索于 2022 年 6 月 16 日，网址：https://www.tiktok.com[15] Xiaoqi Yin、Abhishek Jindal、Vyas Sekar 和 Bruno Sinopoli。2015。基于控制理论的 HTTP 动态自适应视频流媒体方法。《SIGCOMM 计算机通信评论》45 卷第 4 期（2015 年 8 月），325–338。https://doi.org/10.1145/2829988.2787486[16] Chao Yu、Akash Velu、Eugene Vinitsky、Jiax uan Gao、Yu Wang、Alexandre Bayen 和 Yi Wu。2021。近端策略优化在协作式多智能体游戏中的惊人效果。https://doi.org/10.48550/ARXIV.2103.01955[17] Guanghui Zhang、Jie Zhang、Ke Liu、Jing Guo、Jack Lee、Haibo Hu 和 Vaneet Aggarwal。2022。DUASVS：一种短视频中的移动数据节省策略。流媒体。IEEE 服务计算汇刊（2022），1–1。https://doi. org/ 10 1109 2022 3150012 。/TSC。Haodan Zhang、Yixuan Ban、Xinggong [18] Zhang、Zongming Guo、Zhimin Xu、Sheng- bin Meng、Junlin Li 和 Yue Wang。2020。Apl：采用李雅普诺夫优化的短视频自适应预加载。在 2020 IEEE 国际视觉通信与图像处理会议（VCIP）上。IEEE，13–16。[19] Huanhuan Zhang、Anfu Zhou、Yuhan Hu、Chaoyue Li、Guangping Wang、Xinyu Zhang、Huadong Ma、Leilei Wu、Aiyun Chen 和 Changhui Wu。2021。Loki：通过融合基于规则的模型，改善基于学习的实时视频自适应的长尾性能。在第 27 届年度国际移动计算与网络会议论文集（路易斯安那州新奥尔良）（MobiCom '21）中。计算机协会，美国纽约州纽约市，775 788 10 1145 3447993 3483259 –。https://doi.org/ 。/ 。[20] 2016。Jiakai Zhang 和 Kyunghyun Cho。查询高效模仿学习，用于端到端自动驾驶。https://doi.org/ 10 48550 1605 06450 。/ [21] 2022。ARXIV。Chao Zhou、Yixuan Ban、Yangchao Zhao、Liang Guo 和 Bing Yu。PDAS：面向短视频的概率驱动的自适应流媒体。在第 30 届 ACM 国际多媒体会议论文集（葡萄牙里斯本）（MM '22）中。计算机协会，美国纽约州纽约市，7021–7025。https: //doi.org/10.1145/3503161.3551571[22] Xutong Zuo、Yishu Li、Mohan Xu、Wei Tsang Ooi、Jiangchuan Liu、Junchen Jiang、Xinggong Zhang、Kai Zheng 和 Yong Cui。2022。面向短视频流媒体的带宽高效的多视频预取。在第 30 届 ACM 国际多媒体会议论文集（葡萄牙里斯本）（MM '22）中。计算机协会，美国纽约州纽约市，7084–7088。https: //doi.org/10.1145/3503161.3551584。